Machine Learning Evaluation Metrics

Emma Amor
Feb 3 · 8 min read ★

Evaluation metrics help to evaluate the performance of the machine learning model. They are an important step in the training pipeline to validate a model. Before getting deeper into definitions and types of metrics, we need to understand what type of machine learning problem we are solving. Classification metrics differ from regression metrics. These metrics influence how we weight the importance of different characteristics in the results and our ultimate choice of which algorithm/model-version to choose.

Imagine the ants as they are our metrics, and the big rock is our model. They want to make sure that they carried it successfully to their destination.

Classification metrics

- Accuracy
- Precision
- Recall
- F-Score
- ROC (Receiver operating characteristic)
- AUC (Area under the curve)

Let's have a look at the listed metrics above one by one.

Accuracy:

The first metric to evaluate when it comes to a classification problem is the accuracy, which can be calculated using the confusion matrix.

This is how a confusion matrix looks like.

	Predicted: NO	Predicted: YES
Actual: NO	TN = ??	FP = ??
Actual: YES	FN = ??	TP = ??

Confusion matrix

This matrix represents the summary of prediction results which show us the ways in which our model is confused when it makes predictions and it also gives us insights into the types of errors being made(FP/FN*).

- **TP(True positive)**: Observation is positive, and is predicted positive.
- **TN(True negative)**: Observation is negative, and is predicted negative.
- **FP(False positive)**: Observation is negative, and is predicted positive.
- **FN(False negative)**: Observation is positive, and is predicted negative.

Let's imagine that we would like to predict if someone has a heart decease or not. we have 165 patient and we built a model to classify them. The picture below shows our model results.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Heart decease confusion matrix

As we can see we Have 10 patients claimed as they have a heart decease while they don't (FP: Falsely positive). Also, 5 patients are claimed as they do not have a heart decease while they do(FN: Falsely negative).

So, with the help of the confusion matrix, a model's accuracy helps us to find how much we predicted correctly, out of all the classes.

Mathematical Definition

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

For the last example, the model accuracy would be:

acc = 50+100/ (50+10+5+100) = 0,909

Recall or Sensitivity:

The recall is also known as the true positive rate. It is the number of positives our model claims compared to the actual number of positives there are throughout our data.

Mathematical Definition

$$recall = TP / (TP+FN)$$

> *For high recall, we need to minimize FN*

Precision:

Precision is also known as the positive predictive value. It is the number of accurate positives our model claims compared to the number of positives it actually claims.

Mathematical Definition

$$precision = TP / (TP+FP)$$

> *For high Precision, we need to minimize FP*

Recall vs Precision

> **Recall a better measure than precision:**

- Imagine you have thousands of free customers registering on your website every week. The call center team wants to call them all, but it is impossible, so they ask you to select those with good chances to be a buyer. You won't care to call a guy that is not going to buy (so precision is not important) but for you, it is very important that all of them are in your selection(Buyers). That means that my model needs to have a **high recall**, no matter if the precision.

> **Precision a better measure than recall:**

- For recommendation systems like Amazon, Youtube, Netflix or any other, false-negatives are less of a concern. **High Precision** is demanded here.

F-score:

accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost the same.

Let's see this confusion matrix results:

		Predicted/Classified	
		Negative	Positive
Actual	Negative	1001	0
	Positive	1	1

As we can see, from 1003 samples, we have 1002 samples correctly classified. The **accuracy here is equal to 0,999.** hallelujah, you have done a great job! (Now we can celebrate it with some champagne :D !!)

But what if I mentioned that the negative over here is actually someone who is sick and carrying a **coronavirus** that can spread very quickly?!! Well.. you know what will happen :(

Now we realized that accuracy is not the be-all and end-all model metric to use when selecting the best model, we need to use another measure called F-score.

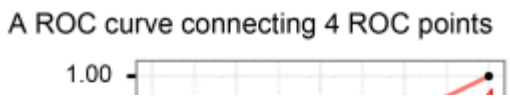
- It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.
- F-score helps to measure Recall and Precision at the same time.
- It is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score.

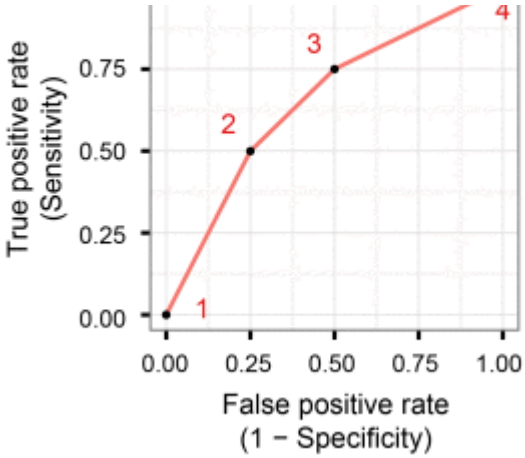
ROC (Reciever operating characteristic curve) :

It is a graph showing the performance of a classification model at all classification thresholds.

This curve represents two axes:

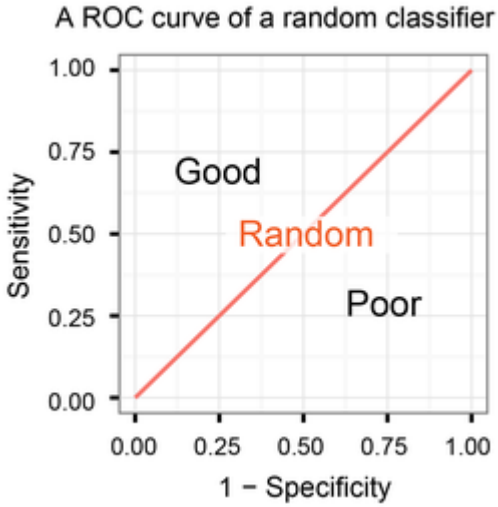
- True positive rate (recall/sensitivity)
- False-positive rate (1- specificity) : FPR = FP/ FP+TN



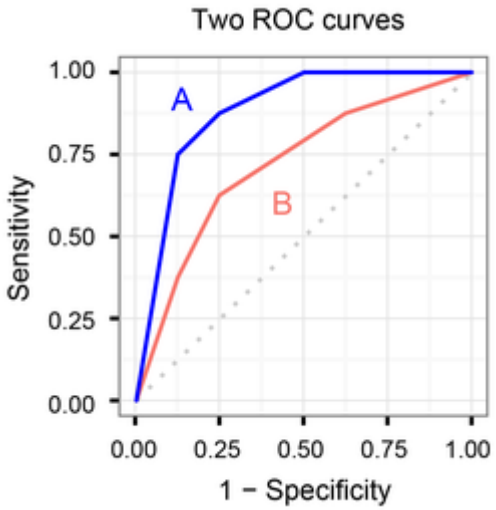


ROC curve

A ROC curve represents a classifier with a random performance level. The curve separates the space into two areas for good and poor performance levels.



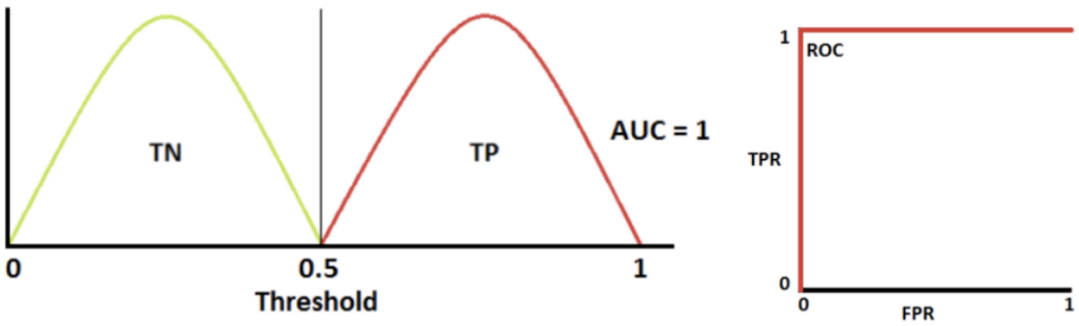
- ROC is usually used to compare classifier performances, Like the figure shows, classifier A outperforms better than classifier B.



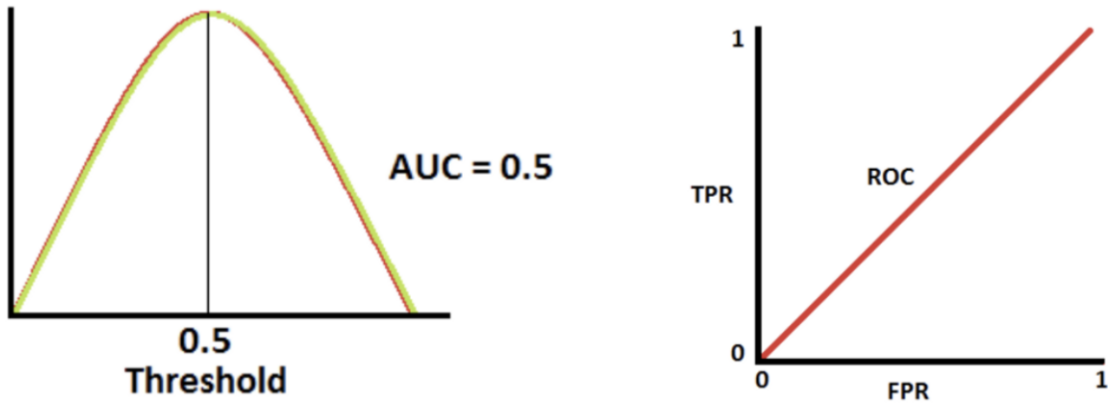
AUC (area under the curve):

It is an area under the curve calculated in the ROC space. It is the metric we consider when we want to evaluate a model's performance when using the ROC curve, also called AUROC. Although the theoretical range of the AUC score is between 0 and 1, the actual scores of meaningful classifiers are greater than 0.5.

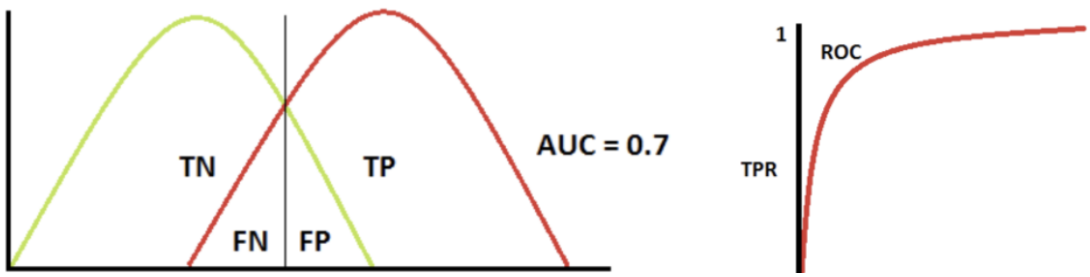
- An excellent model has AUC near to 1, which means it has a good measure of separability.



- A poor model has AUC near to 0, which means it has the worst measure of separability.



- Acceptable models have AUC greater than 0.5.



Regression metrics

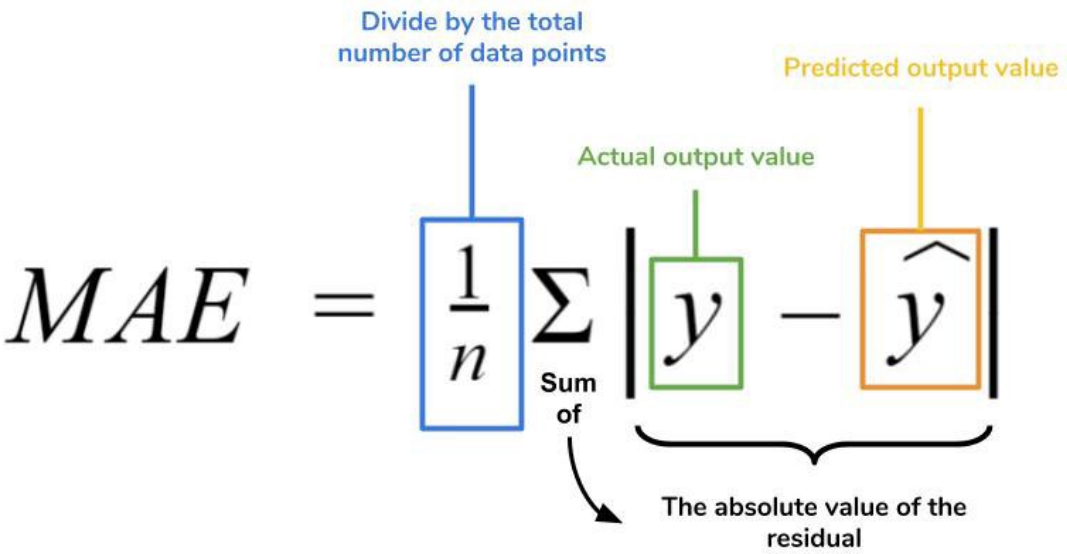
Unlike a classification problem, the objective of a regression problem is not to make predictions on a discrete variable (dog vs cat). Instead, we would be tasked with predicting house prices.

the objective is to evaluate our performance against the ground truth among all the predictions. The ground truth is historical data containing true house prices.

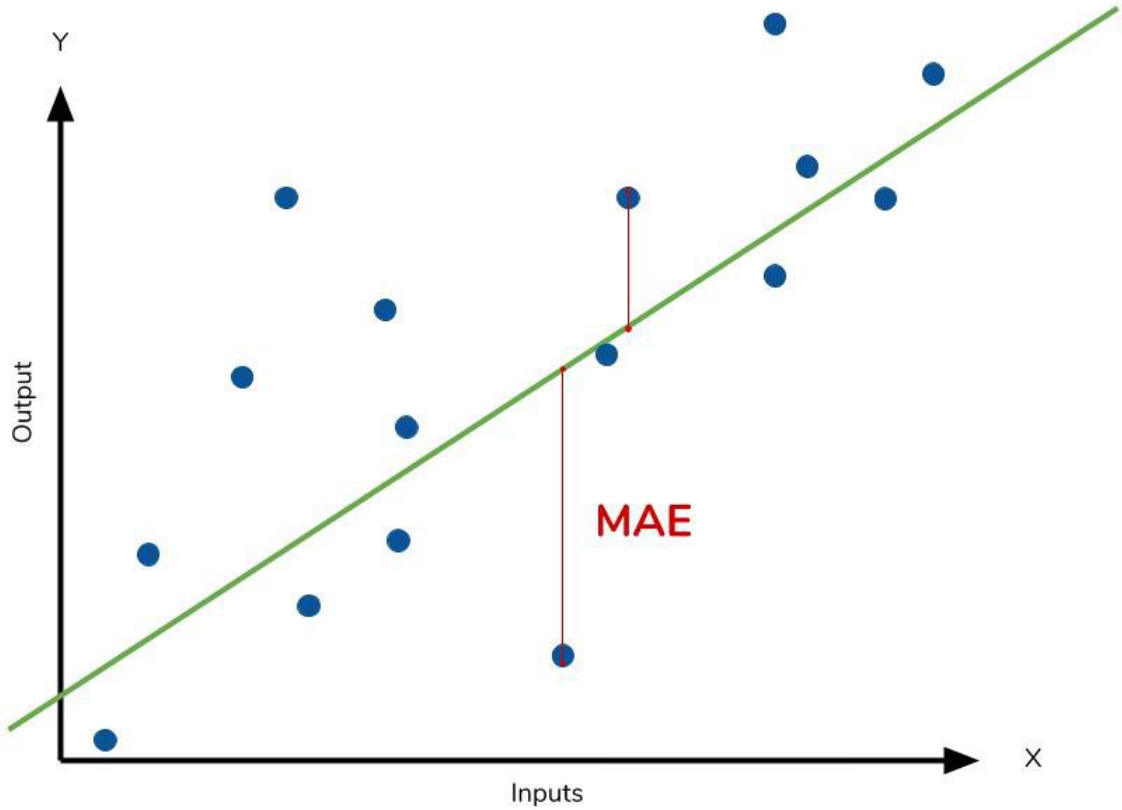
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R2 Score
- Adjusted R2

Mean Absolute Error:

MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight.



It is usually intended to measure average model bias.

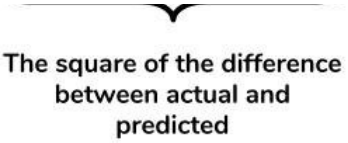


Generally, **the mean absolute error** (**MAE**) is a common measure of forecast error in time series analysis, where the terms “mean absolute deviation” is sometimes used in confusion with the more standard definition of mean absolute deviation.

Mean Squared Error

Both **the mean squared error** and **the mean absolute error** tell you how close a regression line is to a set of points. **MSE** is just like the **MAE** but *squares* the difference before summing them all instead of using the absolute value. We can see this difference in the equation below.

$$MSE = \frac{1}{n}\sum\underbrace{\left(y - \hat{y}\right)^2}$$



as we can conclude from the formula, **MSE** assigns more weight to the bigger errors. The algorithm then continues to add them up and average them. If you are worried about the outliers, this is the metric that you should look at!

R Squared Score

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. It defines the degree to which the variance in the dependent variable (or target) can be explained by the independent variable (features).

Say the R-squared value for our predictive model is 0.8. This means that 80% of the variation in the dependent variable is explained by the independent variables.

the r-squared value would be 1 if the independent variables are able to explain all the variation in the target variable. And this would be the ideal scenario. Thus we can say that higher the r-squared value is, the better is the model.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

SSRES: the Residual sum of squared errors

SSTOT: the total sum of squared errors

The major shortcoming of **R2** is that only the dispersion is quantified if it is used alone. A model that systematically over/under-predicts all the time will still result in good **R2** values close to **ONE** even though all predictions were incorrect. Thus, Generally, it is better to consider ***adjusted R-squared*** rather than **R-squared** according to the nature of your data and model.

Adjusted R2

The **adjusted R-squared** compares the explanatory power of regression models that contain more than one predictor. Suppose you compare a **three-predictor model** (3 features: **Y'i = b0 + b1X1i + b2X2i + b3X3i**) with a higher R-squared to a one-predictor model (**Y'i = b0 + b1X1i**).

We can not compare these 2 models using the **R squared** value. Is the three predictor model better than the one predictor model because it has a higher R squared value? Or, it could be that the R squared is higher because the model has more predictors? **Simply we need to compare the adjusted R-squared values to find out!**

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance (imagine having eyes color as a predictor when trying to predict people with heart decease). The adjusted R-squared can be negative, but it's usually not. It is always lower than the R-squared.

$$Adjusted\ R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Now we know how to evaluate our classification and regression models, we need to make sure that we have an accurate model that is able to generalize on new unseen data. Thus, we need to make sure that the model is not overfitting or underfitting our data.

You can learn more about **overfitting** and **under-fitting**, and how they arise because of **bias** and **variance** of datasets and models here.

Further to learn more about tackling **overfitting** and **under-fitting**, head here

Regularization and tackling overfitting

A cheatsheet to regularization in machine learning

medium.com

Machine Learning Deep Learning Data Science Classification Regression